Cooperation by Design: Leadership, Structure, and Collective Dilemmas

COOPERATION BY DESIGN: LEADERSHIP, STRUCTURE, AND COLLECTIVE DILEMMAS

WILLIAM T. BIANCO
ROBERT H. BATES
Duke University

We return to the analysis of cooperation among interdependent rational individuals. We emphasize the limited impact of iteration (or repeated play) and explore the possibility of an alternative: intervention by rational agents, whom we call leaders. We show that leadership is more significant for initiating cooperation than for sustaining it. In addition, we identify two features of organizations that are critical in determining a leader's ability to initiate and sustain cooperation by structuring the incentives of his followers: the leader's capabilities (information and strategy sets) and reward structure (payoff function).

We examine the role of leadership in "collective dilemmas" —situations where individually rational behavior can prevent the securing of socially rational outcomes (see Axelrod 1981, 1984; Hardin 1982; Olson 1977; Taylor 1987). As many scholars have noted, iteration and retaliatory strategies provide ways to resolve collective dilemmas. Others (Alchian and Demsetz 1972; Frohlich and Oppenheimer 1978; Holstrom 1982; Kreps 1984; Miller 1987, 1988; Popkin 1979; and Putterman 1986) suggest that leaders—players with control over the distribution of benefits generated by collective action—can achieve and enforce cooperation. We provide an analysis of leadership in collective dilemmas.

The following image captures the problem we are examining. A group of rational egoists faces an array of possible equilibria in an iterated dilemma game. Only one of these equilibria yields the full benefits attainable from cooperation; but the players cannot attain it behaving as rational egoists. They therefore designate one of their member as a leader and ask the leader to apportion and withhold benefits in such a way as to make it in their individual self-interest to behave cooperatively. The question is, Can the leader succeed in initiating and sustaining cooperation?[1]

We show that the impact of leadership in collective dilemmas depends on a leader's incentives and capabilities. One might think of these characteristics as features of an institution—powers or resources that accrue to whoever holds the position of leader. Alternatively, one can think of players choosing someone to lead them but also deciding what powers to give to that individual.

Some types of leaders, we show, are able to monitor individual follower strategies and to target sanctions against shirkers. Given an appropriate strategy and "reputation" (follower beliefs about the leaders' payoffs) these leaders can induce other players to begin the game cooperating and continue to do so. However, other types of leaders—those who can

only monitor and sanction groups of players rather than individuals—are generally unable to use their powers to initiate and sustain mutually beneficial cooperation. Thus, a leader's success at "solving the dilemma" depends on the leader's "reputation" and on exogenous features of the game that the leader may not control.

These results underscore an important role for leadership and institutions: by providing targetable and appropriate incentives, they make possible the provision of public goods. As is well known, the production of public goods is bedeviled with inappropriate incentives. By separating consumption from production and by making consumption possibilities contingent on individual effort in production, institutions can provide organizers with the means to motivate self-interested individuals to participate in collective action.

The results also address questions concerning the creation of organizations. Thus, our analysis is relevant to scholars who seek to explain hierarchies in political organizations (Miller 1987; Moe 1984), as well as to scholars who explore the origins of hierarchy and authority relations in market-like settings (e.g., Alchian and Demsetz 1972; Kreps 1984; Williamson 1985). It is relevant, in short, to much of the recent work in the "new institutionalism."

The Problem of Cooperation

The problem of cooperation can be analyzed in the following form. We will specify two iterated games. In the first, the *follower game*, three players, or followers, can engage in mutually beneficial but costly collective action. The second, *leader-follower game*, is the same except that a fourth player, the leader, controls the distribution of benefits produced by collective action. The results of this analysis are easily generalizable to the case of n followers acting by themselves or to the n follower–one leader case. (For an extended discussion of games of this type, see Bianco 1988.)

In the follower game, on each iteration t, follower i's strategy set contains two options: to cooperate ($s_{it} = 1$) or defect ($s_{it} = 0$). However, a follower who cooperates incurs the cost $c > 0$ that the one who defects does not incur. Cooperation also generates benefits that are distributed in equal shares to all followers. Given a vector of follower strategies s_t (one s_{it} per follower), the amount of benefits available for distribution equals $b(s_t)$. The relationship between follower strategies and benefits produced is specified as follows:

$$b(s_t) = a_1(s_{1t} + s_{2t} + s_{3t}) + a_2(s_{1t} \times s_{2t} + s_{1t} \times s_{3t} + s_{2t} \times s_{3t}). \quad (1)$$

When it is possible, we use β_n to refer to the quantity of benefits produced given an s_t in which n followers cooperate. Thus, given an s_t where two followers cooperate, $b(s_t) = \beta_2 = 2(a_1) + a_2$.

It should be noted and stressed that equation 1 is sufficiently general that it captures the characteristics of the major forms of collective dilemmas examined in the literature. For example,

1. Let $a_1 > 0$ and $a_2 = 0$. Then the amount of benefits is a linear function of the number of cooperators. This is the form of collective dilemma analyzed by Hardin (1971, 1982).
2. Let $a_1 \geq 0$ and $a_2 > 0$. Then the function yields team production, a canonical form of production externality that lies at the foundations of much of the new theory of the firm (Alchian and Demsetz 1972).

Either set of conditions yields the classical prisoner's dilemma (Axelrod 1981, 1984; Hardin 1971, 1982; Hardin and

Barry 1982). Given the centrality of the prisoner's dilemma to the literature on cooperation, it may be useful to illustrate the nature of the dilemma faced by the players in this follower game.

The players' payoffs are calculated as follows. A follower i who cooperates on iteration t receives the payoff $v_{it}(s_t) = b(s_t)/3 - c$, while a follower i who defects on iteration t receives the payoff $v_{it}(s_t) = b(s_t)/3$. For example, suppose $a_1 = 2$, $a_2 = 0$, and $c = 1$. Given that all other followers cooperate, follower i receives $(2 \times 3)/3 - 1 = 1$ for cooperating and $(2 \times 2)/3 = 1.33$ for defecting. Similarly, in a game where $a_2 = 2$, $a_1 = 1$, $c = 1.5$ and other players cooperate, follower i receives $[(2 \times 3) + (1 \times 3)]/3 - 1.5 = 1.5$ for cooperating and $[(2 \times 2) + (1 \times 1)]/3 = 1.67$ for defecting.

The dilemma of collective action arises when mutual cooperation makes all followers better off but each follower possesses a dominant strategy of defection at each iteration of the game. As shown previously, β_n is the amount of benefits produced on iteration t when exactly n players cooperate. In this case, a dilemma arises in a collective action game with $a_1 > 0$ and $a_2 = 0$ if and only if

$$\beta_2/3 > c > a_1/3. \qquad (2)$$

If $\beta_2/3 > c$, all followers prefer any outcome where two or more followers cooperate to the outcome where all followers defect.[2] But if $c > a_1/3$, each follower prefers to defect regardless of what other followers do.

Table 1 gives an example of a collective action game satisfying relationship 2: $a_1 = 2$, $a_2 = 0$, and $c = 1$. Consider the leftmost column on Table 1. Assuming that both other followers cooperate, follower i receives $(3 \times 2)/3 - 1 = 1$ for cooperating and $(2 \times 2)/3 = 1.66$ for defecting. Therefore, i is better off defecting, *regardless* what other players do. The same is true if either one or none of follower i's opponents cooperate. However, all followers receive a higher payoff in the two-cooperator outcome (.33) compared with the all-defect outcome (0). Hence, the dilemma.

Table 1. Follower i's Payoff in a Game Where $a_1 = 2$, $a_2 = 0$, and $c = 1$

	Strategies of other followers:		
	Both followers cooperate	One follower cooperates, one defects	Both followers defect
Cooperate $s_{it}=1$	1	.33	-.33
Defect $s_{it}=0$	1.11	.66	0

The same dilemma arises in a team production game when

$$\beta_2/3 > c > [\beta_3 - \beta_2]/3. \qquad (3)$$

In words, $\beta_2/3 > c$ implies that all followers prefer an outcome where two or more followers cooperate to the all-defect outcome where no one cooperates. However, $c > [\beta_3 - \beta_2]/3$ implies that defection is each follower's dominant strategy. The payoff matrix for a team production game satisfying relationship 3—where $a_1 = 2$, $a_2 = 1$, and $c = 1.5$—is given in Table 2. Again, follower i is better off defecting, regardless what other players

Table 2. Follower i's Payoff in a Game Where $a_1 = 2$, $a_2 = 1$, and $c = 1.5$

	Strategies of other followers:		
	Both followers cooperate	One follower cooperates, one defects	Both followers defect
Cooperate $s_{it}=1$	1.5	.16	-.16
Defect $s_{it}=0$	1.66	.66	0

do. However, all followers prefer an outcome where two or more followers cooperate to the all-defect outcome.

A game subject to the constraint defined in equation 1 thus possesses the characteristics of a collective dilemma. In the classic literature on such dilemmas (esp. the prisoners' dilemma) iteration is offered as a way of resolving the problem of cooperation. In the section that follows, we argue that iteration is not enough.

The Necessity for Leadership: Why Iteration Is Not Enough

Robert Axelrod (1981, 1984) demonstrated to political scientists that in situations of repeated play, actors in prisoner's dilemmas might find it to their advantage to choose to cooperate. In more general form, the result is known as the Folk Theorem (see Fudenberg and Maskin 1986 for a review). It has not been sufficiently recognized that the implications of the Folk Theorem are not as optimistic as Axelrod's analysis suggests.

In the iterated version of the follower game, play continues for an infinite number of iterations, beginning with iteration 0. Each follower i will choose a strategy s_i, which gives i's strategy choice on iteration t, s_{it}, as a function of the history of the game, that is, as a function of other follower's strategy choices or of the amount of benefits produced on previous iterations. Follower's payoffs on iterations $t > 0$ are discounted by w^t, where $0 < w < 1$.

The Good News

According to the Folk Theorem, if the followers' discount rates are "high enough," full cooperation can be enforced as a subgame-perfect equilibrium (Selten 1975).[3] The followers need but employ trigger strategies (Friedman 1971; 1986, 85–104), with each follower cooperating until someone defects. The Folk Theorem suggests a focus on the trigger strategy that inflicts the largest possible punishment on a defector—if that threat does not deter defection, no other threat can. This strategy in the follower game will be labeled as "q-trigger" strategy (q for *quantity of benefits*):

q-trigger: $t = 0$: Cooperate.

$t > 0$: Cooperate if $b(s_{t^*}) = \beta_3$ for all $t^* < t$, defect otherwise.

A follower i using q-trigger cooperates as long as everyone else does, but any deviation from cooperation by another follower j triggers "permanent retaliation"—follower i refuses to cooperate on all subsequent iterations.

Contemporary theory indicates that the q-trigger strategy (or any trigger strategy) must be subgame-perfect to sustain cooperation in an iterated game (for accessible discussions of this point, see Bianco 1988; Friedman 1986, pp. 77–82, and 88–92; Ordeshook 1986, 137–42). To say that a situation where followers use the q-trigger strategy is a subgame perfect equilibrium implies two things. First, the follower's threats against would-be defectors are *effective*; a player cannot improve the payoff by deviating from cooperation, given that this deviation "triggers" punishment. Second, follower threats are *credible*; the followers are willing to carry out their threatened punishments if anyone defects.

Proposition 1 (proof in Appendix) gives the Folk Theorem result for the follower game.

PROPOSITION 1. *In a game where followers use the q-trigger strategy, full cooperation can be sustained as a subgame-perfect equilibrium if and only if* $w \geq 3c/\beta_2 - (\beta_3 - \beta_2)/\beta_2$.

The result is what one might expect, given

the work of Axelrod (1981, 1984) and others: cooperation can be sustained as an equilibrium in a game involving repeated play. For example, in the iterated version of the game described in Table 1, the follower's use of the q-trigger strategy will yield full cooperation as an equilibrium outcome if $w \geq .25$. In an iteration version of the Table 2 game, cooperation can be sustained by the q-trigger strategy if $w \geq .1$.

The Bad News

The problem is that the Folk Theorem also states that *any* individually rational outcome can be supported as a subgame-perfect equilibrium, given an appropriate discount rate (Fudenberg and Maskin 1986). Moreover, full defection is a subgame-perfect equilibrium *regardless* of the discount rate.

There is thus a multiplicity of subgame-perfect equilibria in the follower game (and iterated dilemmas in general); full cooperation is but one of them. The problem of securing cooperation, then, is not one of creating incentives for players to cooperate when others are not doing so; nor is it one of *sustaining* cooperation. Rather, the problem of securing cooperation is one of getting to cooperation in the first place, of *attaining* cooperation.

Axelrod and others introduce the mechanism of evolution to resolve this problem (Axelrod 1981, 1984). Cooperation emerges in repeated play because the cooperators displace the noncooperators over time. However, this approach requires players to engage in nonequilibrium (i.e., nonoptimal) behavior. Alternatively, Axelrod posits a major role for norms, showing that societies in which opportunistic behavior is punished become societies inhabited by cooperators (Axelrod 1986). The problem with this argument is that the solution to the problem of cooperation remains exogenous. It is imposed, rather than contrived by those entrapped by the dilemma. We explore the feasibility of an alternative mechanism, that of human design.

Adding a Leader

We now examine the capacity of particular actors, whom we call leaders, to organize incentives to induce followers in a collective dilemma game to move toward mutually beneficial outcomes. In other words, we look at how adding an element of organization to decentralized behavior by interdependent actors might enable them to resolve the problem that repeated play does not resolve, the initiation of cooperation (as opposed to its maintenance).

We now turn to the leader-follower game, in which a new player, the leader (denoted as player 4), controls the distribution of the benefits produced by collective action. In analyzing the impact of leadership, we distinguish between different types of leaders. Leaders can vary in their capabilities and incentives, and these differences matter.

Leader Capabilities

By a leader's capabilities we mean a leader's strategy set and information about the strategies of followers. We distinguish between two levels of such capabilities: enhanced and limited.

An enhanced leader observes each follower's strategy choice s_{it} on each iteration t and can reward or punish each follower separately.[4] Given these capabilities, an enhanced leader is able to "target" rewards and punishments; that is, the leader can give followers the incentive to cooperate by distributing benefits to cooperators and withholding benefits from defectors. A trigger strategy that specifies this behavior will be denoted as

"*e*-trigger" (*e* for *enhanced*). Formally,

e-trigger: $t = 0$: Reward follower i.
$t > 0$: Reward follower i if $s_{it^*} = 1$ on all $t^* < t$, punish follower i otherwise.

Thus, an enhanced leader can reward or punish a follower i regardless of what the leader does to another follower j.

Leaders can also be limited in their capabilities. A limited leader observes only aggregate output, $b(s_t)$, on each iteration t but is unable to discern the followers' individual strategy choices, s_{it}.[5] As a consequence, a limited leader possesses a limited strategy set: on an iteration t, such a leader can either reward or punish the followers but only as a group. In other words, the limited leader can use a trigger strategy, but only against all followers simultaneously. This strategy will be labeled "*l*-trigger" (*l* for *limited*).

l-trigger: $t = 0$: Reward all followers.
$t > 0$: Reward all followers if $b(s_{t^*}) = \beta_3$ for all $t^* < t$, punish all followers otherwise.

Leader Payoffs

The second critical feature is the nature of the leader's payoff function. We investigate two possibilities.

First, the leader could be a residual claimant. A leader who is a residual claimant is compensated for the costs of monitoring and supervision by receiving an initial share of benefits plus all the benefits not distributed to the followers.

In the context of our game, we assume that a residual claimant receives a share equal to σ $(1 > \sigma > 0)$ of each unit of benefits plus all the benefits not distributed to the followers. This characterization produces the following payoff function for a residual claimant leader:

$v_{4t}(s_t) = \sigma b(s_t)$ if leader rewards all followers
$= \sigma b(s_t) + 1/3(1 - \sigma)b(s_t)$ if leader rewards two followers and sanctions one follower
$= \sigma b(s_t) + 2/3(1 - \sigma)b(s_t)$ if leader rewards one follower and sanctions two followers
$= b(s_t)$ if leader sanctions all followers.

Leaders can also be motivated by being paid fixed shares; that is, they would receive a fixed share of the benefits from cooperation but not retain any benefits that they fail to distribute. As with a residual claimant, a fixed-share leader still has an incentive to motivate followers to cooperate: the leader's payoff increases with output. Here we will operationalize this form of leader payoff by assuming the leader always receives the share σ of the benefits produced. In this case, a fixed-share leader receives the payoff $v_{4t}(s_t) = \sigma b(s_t)$, given s_t, regardless of the leader's strategy choice on iteration t.[6]

Motivating leaders by making them residual claimants is a common feature in economic organizations. Alchian and Demsetz (1972) give a classic exposition of the argument. Other examples are found in the recent literature on the firm (Putterman 1986). Similar incentives are imputed to the managers of public bureaucracies, who are held by some to maximize the discretionary portions of their budgets (Bendor, Taylor, and Van Gaalen 1987; Niskanen 1971), that is, the portion over and above the actual costs of public programs.

Examples of leaders motivated by fixed shares would be managers and workers who are paid through profit sharing. In politics, examples would include party bosses (Erie 1988) or committee chairs (Arnold 1979, Ferejohn 1974) who exact a portion of the benefits resulting from collective action, be it the formation of elec-

toral coalitions at the polls or legislative coalitions in the Congress.

The Impact of a Leader: Follower Strategies and Payoffs

We begin by noting how the addition of a leader affects the payoffs and strategies of the followers. The payoff function for follower i on an iteration t of the game becomes

$v_{it}(s_t) = (1 - \sigma)b(s_t)/3 - c$ if follower cooperates and is rewarded
$= (1 - \sigma)b(s_t)/3$ if follower defects and is rewarded
$= -c$ if follower cooperates and is punished
$= 0$ if follower defects and is punished.

As in the follower game, follower i incurs a cost (c) in the leader-follower game whenever i cooperates. However, follower i receives the benefits of the group's efforts only if the leader chooses to reward i. Follower i receives exactly $(1 - \sigma)b(s_t)/3$ units of benefits when the leader rewards i—one-third of the benefits produced, minus the leader's share and any benefits the leader receives as a residual claimant.

It should be noted and stressed that the followers' payoffs under full cooperation with a leader are lower than their payoffs when they achieve cooperation by themselves. In the former case, follower i receives $1/3(1 - \sigma)b(s_t)$ units of benefits, while in the latter i receives $1/3b(s_t)$ units. This reduction in benefits arises because some benefits are used to compensate the leader.

Finally, adding a leader generates implications for the follower's trigger strategies. Without a leader, a follower using the q-trigger strategy cooperates until some other follower defects. However, the leader is now the source of rewards and punishments. In this case, the followers' strategies must be modified to contain threats that deter the leader from making unprovoked punishments. Such threats will be characterized in two ways.

By the first strategy, a follower retaliates if any other follower defects or if the leader fails to give rewards. This strategy is denoted "b-trigger" (b for *both*):

b-trigger: $t = 0$: Cooperate.
$t > 0$: Cooperate if $b(s_{t^*}) = \beta_3$ and the leader rewards follower i on all $t^* < t$, defect otherwise.

A follower i using the b-trigger strategy begins the game cooperating, cooperates as long as the leader rewards i and no follower defects but defects if the leader punishes i or if some follower defects.

Under a second trigger strategy, s-trigger (s for *strategy of leader*), a follower retaliates if the leader fails to reward the follower or if the follower's cost of cooperating exceeds the benefits. Formally,

s-trigger: $t = 0$: Cooperate.
$t > 0$: Cooperate if $b(s_{t^*}) \geq \beta_2$ and the leader rewards follower i on all $t^* < t$, defect otherwise.

We now proceed to the central part of our analysis—the conditions under which cooperation can be sustained and initiated in the leader-follower game.

Sustaining Cooperation with a Leader

We need to characterize the conditions under which cooperation can be sustained in the leader-follower game. To sustain cooperation, the threats of both the leaders and the followers must be effective and credible. Only in this way will coop-

eration be realized as a subgame-perfect equilibrium.

Characterization of the conditions under which cooperation can be sustained in the leader-follower game produces some surprising results. First, the conditions under which cooperation can be sustained with a leader are always narrower than the conditions in a game without a leader (and sometimes knife-edge or nonexistent). Second, these conditions vary with a leader's capabilities and reward structure.

Leaders

Suppose enhanced leaders use the *e*-trigger strategy and limited leaders use the *l*-trigger strategy. Proposition 2 (proof in Appendix) gives the condition under which the leader's retaliatory threats against other followers are effective:

PROPOSITION 2. *A leader's trigger strategy specifies effective threats against defection by follower i if and only if* $w \geq 3c/(1-\sigma)\beta_2 - (\beta_3 - \beta_2)/\beta_2$.

As shown in the Appendix, the effectiveness condition is generated by comparing a follower's payoff under two scenarios: one where the follower plays a trigger strategy (and thus cooperates and is rewarded) and one where the follower switches to an all-defect strategy on iteration t, with retaliation by the leader commencing on iteration $t + 1$. As the value of w increases, the follower's payoff in the first scenario increases, while the payoff in the second remains the same. Thus, leader threats against a follower are effective if the value of w is "high enough." Both leader trigger strategies have the same effectiveness condition because they specify the same retaliation against defectors.

Proposition 3 (proof in Appendix) gives the conditions under which a leader's retaliatory threats are credible:

PROPOSITION 3. *The leader's trigger strategy specifies a credible threat against follower i if and only if a, b, or c is true:*

a. the leader has enhanced capabilities
b. the leader has limited capabilities, is a residual claimant, and $w \geq (1-\sigma)$
c. other followers use the b-trigger strategy.

The significance of Proposition 3 is that the credibility of a leader's retaliatory threats depends on the leader's capabilities and incentives *and* on the strategy used by the followers. Under some conditions, certain leaders will have no incentive to commence retaliation against followers who defect. Such leaders cannot use trigger strategies to sustain cooperation, regardless of the severity of the punishments they could inflict on followers.

The problem of noncredible leader threats arises when the leader's retaliation against a defector "triggers" the remaining followers to switch from cooperation to defection. Obviously, this problem does not arise if followers use the *b*-trigger strategy (and cease to cooperate once someone defects) or if the leader has enhanced capabilities (and can retaliate against one follower while continuing to reward others).

However, consider a leader with limited capabilities, who must punish all followers simultaneously. If the followers use the *s*-trigger strategy, retaliation against a defector will cost a limited leader all future benefits from the remaining followers. Limited leaders who are residual claimants benefit on the initial iteration when they retaliate against a defector, so their threats may be credible. However, fixed-share leaders, whose single-iteration payoff is the same regardless whether they reward or punish followers, need enhanced capabilities in order to have credible threats.

Table 3. Conditions for Sustaining Cooperation, Given Leader Type and Follower Strategies

		Followers' Strategy	
		s-trigger	*b*-trigger
Enhanced Leader	Residual Claimant	$w \geq \max(w_1, w_2)$	
	Fixed Share	$w \geq w_1$	
Limited Leader	Residual Claimant	$w_2 = w \geq w$	$w \geq \max(w_1, w_2)$
	Fixed Share	Never	$w \geq w_1$

$w_1 = 3c/(1-\sigma)\beta_2 - (\beta_3-\beta_2)/\beta_2$
$w_2 = (1-\sigma)$

Followers

Proposition 4 gives the condition under which the followers possess an effective threat against a leader's switch to unprovoked punishment.

PROPOSITION 4. *The followers' trigger strategies (b-trigger or s-trigger) specify effective threats against unprovoked punishment by the leader if and only if a or b is true:*

a. the leader is a residual claimant and $w \geq (1 - \sigma)$

b. the leader receives a fixed share of output.

Again, effectiveness requires that the leader's loss from retaliation equal or exceed the leader's gain from unprovoked punishment. If the leader is a residual claimant, this requirement is met if w is "high enough." Threats against a fixed-share leader are always effective because such a leader does not increase the single-iteration payoff by defecting.

The final step is to specify the conditions under which follower threats against the leader are credible. The following proposition is presented without proof:

PROPOSITION 5. *Follower threats against unprovoked punishments by the leader (as specified by b-trigger or s-trigger) are always credible.*

The intuition behind Proposition 5 is simple: once a leader commences unprovoked punishment of the followers, followers are willing to retaliate because it saves them the cost of cooperation without triggering any reduction in their benefits.

We summarize the information contained in Propositions 2–5 in Table 3, in which we give the conditions for sustaining cooperation in a leader-follower game as a function of leader type (payoff structure and capabilities) and follower strategies.[7] For a specific leader type and follower strategy, each cell in Table 3 records the condition under which the leader has an effective, credible threat against the followers and the followers possess an effective, credible threat against the leader.

The most significant finding contained in Table 3 is that irrespective of the capabilities of the leader or the strategies of the followers, the conditions under which cooperation can be sustained with a leader are narrower than the same conditions in the absence of a leader. In each cell of Table 3, a necessary condition to

sustain cooperation is that

$$w \geq w_1 = 3c/(1 - \sigma)\beta_2 - [\beta_3 - \beta_2]/\beta_2.$$

However, Proposition 1 shows that cooperation can be sustained in a game without a leader if $w \geq 3c/\beta_2 - [\beta_3 - \beta_2]/\beta_2$. Thus, since $1 > \sigma > 0$, adding a leader increases the value of w needed to sustain cooperation. The reason is that some of the benefits of collective action in the leader-follower game are used to compensate the leader. Therefore, the threat of losing all future benefits is a less severe punishment for a would-be defector in the leader-follower game, as compared with the same punishment in the follower game. Since threats in the leader-follower game are less severe, a higher discount rate is required to sustain cooperation.

In addition, Table 3 shows that certain combinations of leader type and follower strategies makes it impossible to sustain cooperation (limited-fixed-share leader, followers use *s*-trigger) or possible only in knife-edge circumstances (residual claimant leader with limited capabilities, followers use *s*-trigger).

In general, adding a leader makes cooperation harder to sustain, not easier. By implication, the benefits of having a leader must lie in the leader's ability to help followers initiate, not sustain, cooperation. We now assess this claim.

The Leader's Role in Initiating Cooperation

To initiate cooperation in a collective dilemma, the leader must somehow give each follower an incentive to cooperate on the *first* iteration of the game, regardless what other followers do. Of course, the follower will prefer to defect if the leader does not have a trigger strategy that generates effective, credible threats against defection. Suppose this requirement is met. In that case, each follower is better off cooperating on the first iteration (and thereafter) if the follower believes that the leader will actually use the trigger strategy when the game is played. That is, the trigger strategy must be the leader's dominant strategy.

As we show here, the first barrier to a leader's success at initiating cooperation is that regardless of leader type, under complete information the trigger strategy is never a dominant strategy. Thus, under complete information a leader cannot use rewards and punishments to initiate cooperation. Of course, as Miller (1987) argues, by transforming the leader-follower game into a game of incomplete information, it is possible to generate follower beliefs (Kreps and Wilson 1982a, 1982b) that a trigger strategy is dominant for the leader regardless of the "true" configuration of leader payoffs. We show here, however, that even the strongest of leader reputations is not sufficient to initiate cooperation. Other things, such as appropriate leader capabilities, are required.

The Dominance of Trigger Strategies

As Proposition 6 shows (proof omitted), under complete information trigger strategies are never dominant strategies for a leader.

PROPOSITION 6. *Regardless of a leader's capabilities and reward structure, a trigger strategy is never the leader's dominant strategy in the leader-follower game.*

This result confirms the conjecture of Miller (1987, 1988) that apart from the effects of reputation, a leader's promise to reward and punish according to a trigger strategy is not credible. Put simply, it is always possible to devise a follower strategy such that the leader prefers to deviate from the trigger strategy.[8] For example, suppose followers use the following strategy:

$t = 0$: Defect.

$t > 0$: Cooperate if the leader punished on iteration 0, defect otherwise.

If followers use this strategy, a leader (regardless of type) receives a higher payoff by punishing on the first iteration and rewarding thereafter than by using a trigger strategy (which will trigger defection by the followers beginning on iteration 1). The trigger strategy cannot, therefore, be a dominant strategy for the leader.

Some types of leaders will prefer to deviate from their trigger strategies even when followers use trigger strategies or the all-defect strategy. If a residual claimant leader expects a follower to defect on iteration t (including $t = 0$), the leader can improve the payoff by punishing the defector on iteration t instead of waiting to commence punishment on iteration $t + 1$ as the trigger strategy specifies. If followers use the *b*-trigger strategy, residual claimant leaders with enhanced capabilities can improve their payoff, given that one follower defects, by deviating from the *e*-trigger strategy and punishing all followers after the defection. In both cases, the leader's incentive to deviate from the trigger strategy implies that the strategy is not the dominant strategy.

Fixed-share leaders face similar problems. A fixed-share leader with limited capabilities is not willing to retaliate against a defector unless other followers use the *b*-trigger strategy. This result is detailed in Proposition 3. Again, the implication is clear: the leader's trigger strategy is not the dominant strategy.

The point is simple: under complete information, even if leaders and followers possess effective, credible retaliatory threats, a leader's trigger strategy is never the dominant strategy. No type of leader is willing to reward followers initially, continue to reward cooperation, and punish defection regardless of the follower's strategies and other parameters in the game. Thus, under complete information leaders face the same problem as followers in a game without a leader—they are willing to use a trigger strategy, but only if everyone else does. Equivalently, the Folk Theorem has the same implications for the leader-follower game as for the follower game: given a high enough discount rate, there are infinitely many equilibria—including full cooperation and full defection. Adding a leader does not reduce the number of equilibria or drive followers to the full cooperation outcome.

Adding Reputations

Suppose the followers are *certain* (because of their beliefs about the leader's payoffs) that the leader will reward and punish according to a trigger strategy (*l*-trigger for a limited leader, *e*-trigger for an enhanced leader). Assume, further, that the leader's threats are effective; this, along with the follower's beliefs, ensures that the leader's trigger strategy will sustain cooperation. Then, to initiate cooperation, given followers' beliefs, a strategy which specifies cooperation initially and as long as the leader rewards must maximize the payoff of each follower. (Here we focus on the strategies *b*-trigger and *s*-trigger.) If some other strategy yields a higher payoff for the follower in some situation, reputation effects will not serve to initiate cooperation.

This argument provides a simple test of the ability of reputation to induce cooperation. Proposition 7 (proof in Appendix) summarizes the result of this test:

PROPOSITION 7. *If followers are certain that the leader will use the strategy l-trigger if the leader has limited capabilities and e-trigger if the leader has enhanced capabilities and in addition* $w \geq w_1$, then

a. if the leader has enhanced capabilities, b-trigger or s-trigger maximizes a follower's payoff

b. if the leader has limited capabilities, a follower's payoff-maximizing strategy depends on the strategies used by other followers.

Thus, the follower's response to a leader's "reputation" of rewarding cooperation and punishing defection depends on the leader's capabilities. If the leader possesses enhanced capabilities, the follower is best off cooperating on the first iteration and thereafter—given beliefs, the follower expects this behavior to be rewarded regardless of what the other followers do. However, if the leader has limited capabilities, a follower will be punished if *anyone* defects. Thus, the follower is willing to cooperate initially only if everyone else does so. In summary, reputations and effective threats serve to initiate and sustain cooperation, given enhanced leaders but not given leaders with limited capabilities.

Conclusion

Employing a game that highlights the dilemmas confronting interdependent rational individuals, we exposed the limited impact of iteration (or repeated play) and explored the possibility of an alternative —rational intervention by a third party, whom we have called a leader.

Our analysis has shown that

1. In collective dilemmas, leadership is more significant for initiating cooperation than for sustaining it.
2. Leaders must be differentiated in terms of capabilities (information and strategy sets) and incentives (payoff functions).
3. Variation in the capabilities and reward structures for leaders strongly condition their ability to sustain and to initiate cooperation in collective dilemmas.

Rather than positing an enlightened leader who intervenes in a welfare-enhancing manner, we have instead made the behavior of the leader endogenous. An important feature of our analysis is that it offers insight into the impact of organization. The features of the organization that structure the choices of leaders are the system of rewards and the information and incentives at the leader's command. Our analysis suggests that proponents of cooperation would be wise to endow their leader with enhanced capabilities. Given effective threats and an appropriate reputation, a leader with enhanced capabilities can use a trigger strategy to initiate and sustain cooperative behavior by followers. The effect of a leader's reward structure on the ability to initiate cooperation is less clear. Regardless of how a leader is compensated, it is possible to construct follower beliefs about leader payoffs such that the followers will expect the leader to reward and punish according to a trigger strategy when the game is played. However, rewarding a leader with a fixed share appears to simplify the problem of constructing a reputation that convinces followers that the leader will use a trigger strategy. Fixed-share leaders appear willing to reward followers initially and on subsequent iterations unless the rewards induce follower defections. In contrast, even if rewards do not induce defections, residual claimant leaders may have an incentive to punish followers initially or to punish spontaneously on subsequent iterations.

Appendix: Proofs of Propositions

Proof of Proposition 1. Suppose all followers use the q-trigger strategy. In this case, follower i's payoff for the multiple iteration game equals $(\beta_3/3 - c)/(1 - w)$. If follower i makes the payoff-maximizing deviation from full cooperation (i.e., to

full defection) and other followers respond by ceasing to cooperate, i's payoff for the whole game equals $\beta_2/3$. Note that each follower j's ($j \neq i$) response to follower i's defection is optimal, given that all followers so respond. In this case, follower i lacks incentive to switch strategies if and only if $(\beta_3/3 - c)/(1 - w) \geq \beta_2/3$ or $w \geq 3c/\beta_2 - (\beta_3 - \beta_2)/\beta_2$. Thus, the vector where all followers use the q-trigger strategy is subgame-perfect if and only if this condition holds. QED

Proof of Proposition 2. Consider a follower i. If i (along with the other followers) uses a trigger strategy, i's payoff for the entire game equals $[(1 - \sigma)\beta_3/3 - c]/(1 - w)$. If follower i defects and the leader retaliates, follower i's payoff equals $(1 - \sigma)\beta_2/3$. Thus, the leader's threat is effective if and only if $[(1 - \sigma)\beta_3/3 - c]/(1 - w) \geq (1 - \sigma)\beta_2$ or $w \geq 3c/(1 - \sigma)\beta_2 - (\beta_3 - \beta_2)/\beta_2$. QED

Proof of Proposition 3. Consider a game where a follower i switches to full defection beginning on iteration t. The leader's threat (regardless of whether the leader uses l-trigger or e-trigger) is to cease rewarding follower i beginning on iteration $t + 1$.

If the remaining followers use b-trigger, the leader's threat is always credible: no follower will cooperate after follower i defects, thus the leader does not lose any future benefits from punishing follower i.

The situation is different when the remaining followers use s-trigger. A leader with enhanced capabilities can punish follower i without provoking any other followers to defect, thus the leader's threat is credible.

If the leader has limited capabilities and followers use s-trigger, retaliation against follower i provokes the other followers to cease cooperating beginning on iteration $t + 2$. If the leader is also a residual claimant, the leader's payoff from iteration $t + 1$ forward, given that the leader uses l-trigger (and punishes everyone beginning on $t + 1$) equals β_2, while the payoff from not retaliating equals $\sigma\beta_2/(1 - w)$. Thus, the payoff from retaliating equals or exceeds the payoff from not retaliating (and thus the threat is credible) if and only if $\beta_2 \geq \sigma\beta_2/(1 - w)$ or $(1 - \sigma) \leq w$. Finally, if the limited leader receives a fixed share and followers use s-trigger, the payoff from retaliating equals $\sigma\beta_2$ and from not retaliating equals $\sigma\beta_2/(1 - w)$. Since this leader's payoff from retaliation is always lower, the threat is never credible.

In summary, the leader always has a credible threat if followers use b-trigger; if the leader has enhanced capabilities; or if the leader has limited capabilities, is a residual claimant, and $w \leq (1 - \sigma)$. The leader's threat is never credible when the leader has limited capabilities and receives a fixed share of output and followers use the s-trigger strategy. QED

Proof of Proposition 4. Suppose a leader contemplates switching to the payoff-maximizing defection strategy (all-defect) on iteration t.

If the leader is a residual claimant, the payoff from t forward, given that the leader switches and followers retaliate on $t + 1$ (follower retaliation is the same regardless of whether they use b-trigger or s-trigger), equals $\beta_3 + 0$. However, if the leader continues to use a trigger strategy, the payoff from iteration t forward equals $\sigma\beta_3/(1 - w)$. Thus, the follower's threat is effective if $\sigma\beta_3/(1 - w) \geq \beta_3$ or $w \geq (1 - \sigma)$.

If the leader receives a fixed share, the payoff from all-defect equals $\sigma\beta_3$, while the payoff from using a trigger strategy equals $\sigma\beta_3/(1 - w)$. Thus, the follower's threat is always effective in this case. QED

Proof of Proposition 7. If the leader is enhanced and uses e-trigger, follower i will be rewarded until i defects. By Proposition 2, follower i lacks incentive to switch from b-trigger or s-trigger to any defection strategy if and only if $w \leq 3c/(1 - \sigma)\beta_2 - (\beta_3 - \beta_2)/\beta_2 = w_1$. Thus,

either trigger strategy maximizes follower i's payoff.

If the leader is limited and uses l-trigger, follower i again has no incentive to provoke retaliation by being the first to defect, given $w \geq w_1$. However, if some other follower defects on iteration t, follower i (along with everyone else) will be punished beginning on iteration $t + 1$. In this case, follower i is better off defecting on iteration t rather than using the trigger strategy (defecting on iteration $t + 2$ and after for s-trigger and on $t + 1$ and after for b-trigger). Thus, a trigger strategy maximizes follower i's payoff, given a limited leader and $w \geq w_1$ only if all other followers use trigger strategies. However, if some follower uses a strategy that specifies defection beginning on iteration t, follower i maximizes the payoff by using the same strategy. QED

Notes

We would like to thank Randy Calvert, William Keech, Robert Keohane, Peter Lange, Terry Moe, and especially Gary Miller for helpful comments. Earlier versions of this paper were presented at the 1988 annual meeting of the American Political Science Association, Chicago, and circulated as Duke University Political Economy Working Paper No. 48. This research was supported by a grant from the Ford Foundation and from the National Science Foundation (grant number SBS-8821151).

1. We focus on the question of whether control over benefits gives the leader any ability to initiate and sustain cooperation in iterated dilemma games. Our analysis does not consider the role of leaders in games that model other situations of cooperation under anarchy, such as chicken, stag hunt, battle of the sexes, etc. For insightful discussions of these games, see Oye 1985 and Schelling 1960. Our analysis also ignores the fact that leaders may also help to achieve cooperation by helping followers coordinate their actions or by persuading followers to change their feelings and beliefs. For discussions of coordination, see Banks and Calvert 1989 and Calvert 1989. The cognitive and emotional aspects of leadership are investigated more fully in the sociology and psychology literature and remain a challenge to rational choice analyses of leadership.

2. A weaker condition is possible: the "dilemma" would still arise if all followers (rather than two) had to cooperate to produce an outcome preferable to all-defect. The condition specified here allows examination of certain follower-leader interactions, which will be described.

3. We focus on the follower's (and later the leader's) ability to enforce full cooperation with retaliatory threats. There are three reasons to focus on full cooperation: the outcome maximizes the payoffs of cooperation, maximizes the severity of threats against would-be defectors, and is arguably a "Schelling point" (Schelling 1960).

4. Such a leader is specified by Alchian and Demsetz (1974, 118) and Miller (1988, 13).

5. This type of leader is developed by Holmstrom (1982), and mentioned by Miller (1988). This construction is consistent with the assumption that the leader can in principle observe follower strategies but that it is prohibitively costly to do so.

6. This specification of a fixed-share leader *does not* satisfy the "budget balancing" condition given in Holmstrom 1982. Intuitively, if a fixed-share leader punishes the workers, all residual benefits above the leader's share are not distributed to anyone. As Holmstrom's critics point out (Eswaran and Kotwal 1984), if these benefits are given to an additional player (a "sponge"), this additional player may have an incentive to disrupt successful team production and thereby increase personal payoff. Here we focus on the effect of having a fixed-share leader without considering how undistributed benefits are allocated.

7. Consider the case where the leader is a limited residual claimant and the other players use the s-trigger strategy. The leader's threats are effective if and only if $w \geq w_1$ (Proposition 2) and credible if and only if $w \leq w_2$ (Proposition 3). Player threats against the leader are effective if and only if $w \geq w_2$ (Proposition 4), and always credible (Proposition 5). Putting these requirements together, the leader and the players have effective, credible threats if and only if $w_2 = w \geq w_1$.

A second example. Suppose the leader is an enhanced residual claimant and the other players use the b-trigger strategy. The leader's threats are effective if and only if $w \geq w_1$ (Proposition 2) and always credible (Proposition 3), while the players' threats are effective if and only if $w \geq w_2$ (Proposition 4) and always credible (Proposition 5). Thus, cooperation can be sustained in this case if and only if $w \geq \max(w_1, w_2)$.

8. We thank Randy Calvert, University of Rochester, for his help in clarifying this point.

References

Alchian, Armen, and Harold Demsetz. 1972. "Production, Information Costs, and Economic Organization." *American Economic Review* 62: 777–95.

Arnold, R. Douglas. 1979. *Congress and the Bureaucracy*. New Haven: Yale University Press.

Axelrod, Robert. 1981. "The Emergence of Cooperation among Egoists." *American Political Science Review* 75:306–18.

Axelrod, Robert. 1984. *The Evolution of Cooperation.* New York: Basic Books.

Axelrod, Robert. 1986. "An Evolutionary Approach to Norms." *American Political Science Review* 80:1095–1112.

Banks, Jeffrey S., and Randall Calvert. 1989. "Communication and Efficiency in Coordination Games." University of Rochester. Typescript.

Bendor, Jonathan, Serge Taylor, and Roland Van Gaalen. 1987. "Politicians, Bureaucrats, and Asymmetric Information." *American Journal of Political Science* 31:796–828.

Bianco, William T. 1988. "The Limits of Cooperation: Sanctioning Problems in Dilemma Games. Presented at the annual meeting of the Midwest Political Science Association, Chicago.

Calvert, Randall. 1989. "Coordination and Power: The Foundation of Leadership among Rational Legislators." University of Rochester. Typescript.

Erie, Steven P. 1988. *Rainbow's End.* Berkeley and Los Angeles: University of California Press.

Eswaran, Mukesh, and Ashok Kotwal. 1984. "The Moral Hazard of Budget Breaking." *Rand Journal of Economics* 15:578–81.

Ferejohn, John A. 1974. *Pork Barrel Politics.* Stanford: Stanford University Press.

Friedman, James W. 1986. *Game Theory with Applications to Economics.* New York: Oxford.

Friedman, James W. 1971. "A Non-Cooperative Equilibrium for Supergames." *Review of Economic Studies* 38:1–12.

Frohlich, Norman, and Joe A. Oppenheimer. 1978. *Modern Political Economy.* Englewood Cliffs, NJ: Prentice-Hall.

Fudenberg, Drew, and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54:533–44.

Hardin, Russell. 1982. *Collective Action.* Baltimore: Johns Hopkins University Press.

Hardin, Russell. 1971. "Collective Action As an Agreeable *n*-Prisoners' Dilemma." *Behavioral Science* 16:472–81.

Hardin, Russell, and Brian Barry. 1982. *Rational Man and Irrational Society.* Beverly Hills, CA: Sage.

Holmstrom, Bengt. 1982. "Moral Hazard in Teams." *Bell Journal of Economics* 13:324–40.

Kreps, David. 1984. "Corporate Culture and Economic Theory." Stanford University. Typescript.

Kreps, David M., and Robert Wilson. 1982a. "Sequential Equilibria." *Econometrica* 50:863–90.

Kreps, David M., and Robert Wilson. 1982b. "Reputation and Imperfect Information." *Journal of Economic Theory* 27:253–79.

Miller, Gary. 1987. "Administrative Dilemmas: The Role of Political Leadership." Washington University. Typescript.

Miller, Gary. 1988. "Individual Rationality in Hierarchies." Washington University. Typescript.

Moe, Terry. 1984. "The New Economics of Organization." *American Journal of Political Science* 28:738–77.

Niskanen, William A. 1971. *Bureaucracy and Representative Government.* Chicago: Aldine-Atherton.

Olson, Mancur. 1977. *The Logic of Collective Action.* Cambridge: Harvard University Press.

Ordeshook, Peter. 1986. *Game Theory and Political Theory.* Cambridge: Cambridge University Press.

Oye, Kenneth A. 1985. "Explaining Cooperation under Anarchy: Hypotheses and Strategies." *World Politics* 38:1–24.

Popkin, Samuel P. 1979. *The Rational Peasant.* Berkeley: University of California Press.

Putterman, Louis, ed. 1986. *The Economic Nature of the Firm.* Cambridge: Cambridge University Press.

Schelling, Thomas C. 1960. *The Strategy of Conflict.* Cambridge: Harvard University Press.

Selten, Reinhard. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory* 4:25–55.

Taylor, Michael. 1987. *The Possibility of Cooperation.* Cambridge: Cambridge University Press.

Williamson, Oliver. 1985. *The Economic Institutions of Capitalism.* New York: Free Press.

William T. Bianco is Assistant Professor and Robert H. Bates is Henry Luce Professor, Department of Political Science, Duke University, Durham, NC 27706.